Rafeeq Shodeinde

Unix Tools project

5/2/2019

Using Unix Tools to web scrape

For this project, my goal was to parse through an HTML file to extract relevant data about premier league football match statistics from the HTML table in the file. The data will be used for performing Machine Learning analysis. For simplicity as well as to avoid repetition, this was done on only one set of data instance. The make file created can be applied to similar Wikipedia pages.

The data was extracted by scrapping individual columns of the table using “egrep” and “sed” then saving them to individual csv files. After all important columns are extracted, the Unix tool “paste” command is used to combine the individual csv files to have a complete table. To make scrapping easier all character before the start of the table and after the end of the table are removed as well as all remaining HTML tags. The data acquired at the end of the end of the process is a complete premier league table in order of highest rank to the lowest rank.

How it works:

1. Download the wiki page to be scrapped: https://en.wikipedia.org/wiki/2016%E2%80%9317_Premier_League.
2. Bash scripts are created in the make file to create individual columns of table which are pieced together to form a whole complete table.